DOCUMENT RESUME

ED 081 589 SE 016 527

TITLE	Improving the Evaluation of Peace Corps Training Activities. Volume II of the Report of Supplemental Activities.
INSTITUTION	Center for Research and Education, Denver, Colo.
SPONS AGENCY	ACTION, Washington, D. C.
REPORT NO	PC-72-42043
PUB DATE	4 Jun 73
NOTE	75p.
AVAILABLE FROM	Center for Research and Education, 2010 East 17th Avenue, Denver, Colorado 80206 ($3.00 plus postage)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Cross Cultural Training; Evaluation; *Measurement Instruments; *Program Effectiveness; *Rating Scales; *Systems Analysis
IDENTIFIERS	*Peace Corps; Research Reports

ABSTRACT

The purpose of this project/study was to review the evaluation system presently being used to assess the effectiveness of Peace Corps training activities in Brazil and to modify rating instruments and scales in order to obtain more accurate measurement. This volume, the second of two reports resulting from the project, describes the procedures and methods used, followed by the scaling instruments produced, complete with detailed instructions for their use, scoring, and accuracy computation. Evaluation scaling systems are developed for both training programs and training activities. (For related document, see Volume I, SE 016 526.) (BL)

ED 081589

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IMPROVING THE EVALUATION

OF

PEACE CORPS TRAINING ACTIVITIES

Volume II
of the report of supplemental activities
conducted under ACTION Contract PC-72-42043

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Leonard Kata
CRE
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

© 1973
Center for Research and Education
Denver, Colorado (USA)

June 4, 1973

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Joseph F. Radford
ACTION
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

SE 016 527

CENTER FOR RESEARCH AND EDUCATION — 2010 E. 17TH AVENUE — DENVER, COLORADO 80206
PHONE (303) 388-6311 CABLE: CREDEN

This material may not be reproduced in whole or in part without the written permission from the Center for Research and Education and ACTION.

Copies of this publication are available for the cost of reproduction and handling ($3.00 plus postage) from the Center for Research and Education, 2010 East 17th Avenue, Denver, Colorado 80206.

VOLUME II

Table of Contents

		Page
Preface		
CHAPTER I.	INTRODUCTION	1
CHAPTER II.	PROCEDURES AND METHODS	
	The Retranslation Method	3
	The Mixed Standard Scale	4
	Development of Training Program Evaluation Scales	9
	Development of Training Activity Evaluation Scales	19
CHAPTER III.	PRODUCTS	
	Training Program Evaluation Scales	25
	Training Program Evaluation Scales	29
	Scoring Keys	34
	Scoring Matrix	36
	Scoring Work Sheet	37
	Summary Sheet	38
	Error Computation Work Sheet	39
	Error Key	42
	Example of Error Computation	43
	Training Activity Evaluation Scales	46
	Training Activity Evaluation Scales	49
	Scoring Keys	52
	Scoring Matrix	57
	Scoring Work Sheets	58
	Summary Sheet	63
	Error Computation Work Sheets	64
	Error Key	65
CHAPTER IV.	CONCLUSION	67
List of References		69

Preface

(We are repeating here the preface to Volume I of the report of supplemental activities conducted under Peace Corps Brazil Training Contract PC-72-42043 to give you the context in which to read Volume II which is contained in this document.)

The study described here was performed during the two-month period of January-February, 1973. A simple statement like this seems rather meaningless apart from the full realization of the tremendous complexity of the study and the scope of tasks involved. The only other study conducted in Brazil with any similarity to this one was the Sao Francisco Valley Evaluation Project completed by Wayne Holtzman and associates of the University of Texas in 1966. The outcomes of that three-year effort, compared to those of the present two-month study, give some perspective to what we were able to accomplish in such a short period of time.

In addition to the constant pressure of time, the large distances involved and the accompanying logistical problems were the major difficulties encountered in completing the work. Maintaining a tight discipline in the rigorous implementation of the study design was difficult, to say the least, when operating from a Colorado base, through our Brazil office, and from there covering a major portion of the large expanses of Brazil. The success we were able to achieve is due to the untiring efforts of a very talented staff and the impressive cooperation of the Peace Corps staff and Volunteers in Brazil.

The project was originally designed according to three different tracks, or intended outcomes:

A. The design of a system for measuring cross-cultural learning and change,

B. The design of evaluation instruments and procedures to accurately assess the effectiveness of specific training activities, and

C. Recommendation of improvements in assessing cross-cultural training needs and improvements in training by establishing those benchmark requirements of cross-cultural experience which should be incorporated into training.

As the project got under way, it soon became apparent that Tracks A and C were so interrelated and dependent upon one another that they should be combined, while Track B could be accomplished somewhat independently of the other two. Accordingly, the project consists of two components: One we titled "Improving Cross-Cultural Training and the Measurement of Cross-Cultural Learning" (Tracks A and C) and the other, "Improving the Evaluation of Peace Corps Training Activities" (Track B).

The first component has been written under this cover as Volume I. The second component has been written under separate cover as Volume II.

The study was coordinated by Dr. Michael F. Tucker, Associate Director of the Center for Research and Education. Project staff for the first component included Howard A. Raik, David L. Rossiter, and Dr. Michael J. Uhes. Paul C. Jorgensen, a CRE/Brazil staff member, participated in the field work in Brazil and during the early drafting phase in Denver. Mr. Raik completed a significant amount of the final drafting of the manuscript.

Mrs. Joanna D. Garver was the project staff member primarily responsible for the second component.

Thomas Brand, Allan S. Dorsey and Guaraciema Rodriguez Dorsey, CRE/Brazil staff members, assisted in the final preparation and translation of the data collection instruments and in the interviewing of Volunteers in the field. Jose Eduardo Barbosa provided logistic and administrative support from the CRE office in Belo Horizonte.

Delano M. B. Carvalho, Edmar da Costa Marques, Jose Afoneo de Melo, and Adalberto Ribeirò collaborated by interviewing Brazilian associates of

the Volunteers during the interview phase. Paulo Assis identified the Brazilian sample and completed the data collection for this group.

Drs. Charles Wagley, T. Lynn Smith and Maxine Margolis of the University of Florida and Dr. Harry W. Hutchinson of the University of Miami aided in conceptualizing the theoretical basis for the Cultural Dimensions Test. Dr. Margolis also prepared test items for inclusion in the Factual Information Test.

Dr. Daniel Anderson, Gary Hodson and Sandy Hodson of the University of Northern Colorado identified the sample of Americans with no Brazil experience (called "Naive Americans" for the purposes of this report), collected data from this group, and performed the statistical analyses on all the collected data.

James Doxsey provided consulting assistance in the early phases, particularly with respect to the measurement of affective behavior. Dr. Lawrence R. James provided consulting assistance in psychometrics and scaling, and was instrumental in outlining the scaling procedures followed to produce instruments for training activity evaluation.

Associate Brazil Peace Corps Respresentatives Vitor Braga (Minas Gerais), Charles Cox (Ceara), James LaFleur (Bahia), Cornelio Lana, Denis L. Lynch (Mato Grosso), Marco Mota (Pernambuco), and George Van Antwerp (Rio Grande do Norte) and Program and Training Officer Robert Gentile provided nominations for Volunteer samples and invaluable logistic assistance.

We wish to thank the Volunteers interviewed in Minas Gerais, Bahia, Mato Grosso, Pernambuco, Rio Grande do Norte, and Ceara, the personnel of the agencies with which they are working, and their associates whom we interviewed, for their patience and cooperation.

We wish to thank also the men and women in Minas Gerais and in Colorado who volunteered to be tested for the Brazilian and Naive American samples.

We feel that the benefits derived from a project of this nature are extremely important to the improvement of Peace Corps training. This study is the first opportunity of this type CRE has had in the three years since completing the Guidelines for Cross-Cultural Training. Training for cultural adaptation is a complicated matter, the very nature of which undergoes rapid change as understanding develops through experience. We hope that what we have accomplished will be of use to Peace Corps trainers in pushing forward the "state of the art" and in helping Americans adapt to other cultures in more effective ways. As is true of most endeavors of this sort, this study is just a start; we hope this effort will provide a base for continuation, modification, and improvement of cross-cultural training and measurement.

Michael F. Tucker
Center for Research and Education

Denver, Colorado
March 1973

CHAPTER I. INTRODUCTION

The purpose of this project was to review the evaluation system presently being used to assess the effectiveness of Peace Corps training activities in Brazil and to modify rating instruments and scales in order to obtain more accurate measurements.

The present system, often referred to as the "weekly" or "running" evaluation of training, was instituted in Brazil training programs by the Peace Corps in order to provide a consistent flow of information regarding the progress of training operations. It consists of three questions which ask trainees to respond, according to a seven-point scale, as to the effectiveness of various training activities. The questions have to do with (1) how well a particular training activity was conducted, (2) how important and significant the content of the activity was, and (3) how relevant the learning achieved was to the future Volunteer job.

The primary weaknesses in this system can be listed as follows:

- The three questions do not represent a good composite of training activity assessment. Such things as the objectives of the activity and quality of materials used are not included.
- The three questions are not unidimensional. In fact, they probably represent seven or more dimensions. The first question asks for a single response to how well the activity was conducted; how good the trainer was in conducting the activity; and how good the method of training was. The second question asks for a single response to how important and how significant the content of the activity was. The third question asks for a single response to how relevant the learning achieved was to the future Volunteer job, and also calls for a judgment of whether something was

learned that changed behavior, would affect future performance on the job, and would be applied on the job.

- These three questions were developed and written by Peace Corps and training staff professionals, so that their intent, meaning, importance, relevance and "face validity" is not always clear to trainees.

- The seven-point rating scales were not well developed and are not psychometrically sound. No reliability or validity data exists for these scales, but several indications of imprecision are apparent. Each point on the scale is not psychologically anchored, i.e., there is no provision for assessing consistency in meaning among trainees for giving a "3" rating versus a "5" (for example) on any given judgment. The respondent simply "rates toward 7" or "rates toward 1" for each question.

  Data resulting from these scales suffer to a great extent from the two most common errors of rating -- the halo effect and leniency. The halo effect means that a given training activity is rated the same across the three different questions due to a generalized feeling about the activity rather than the intended, three-criterion discriminatory judgments. Leniency errors result in rating distributions that are inaccurately high and are therefore skewed and have small ranges, i.e., the mean for a seven-point scale should generally be 4.0 with a normal distribution, but the means for these scales are almost always much higher with skewed distributions.

This project was undertaken to develop scales that would overcome these weaknesses and result in an evaluation method that would accurately reflect the quality of training activities. The remainder of this report will describe the procedures and methods used, followed by the scaling instruments produced, complete with detailed instructions for their use, scoring, and accuracy computation.

CHAPTER II. PROCEDURES AND METHODS

The first task was to review the literature on evaluation strategies, psychometric theory, and scaling methods (see the list of references on Page 69) with the objective of selecting an approach or a combination of approaches that would overcome each of the deficiencies in the present assessment procedure. Two methods emerged that appeared especially promising, both of which had been developed for use in rating personnel performance. The first was developed over ten years ago (Smith and Kendall, 1963) and is called the Retranslation of Expectations method of constructing unambiguous anchors for rating scales. This method was selected for the purpose of constructing properly worded statements for the scales. The second method is a recent innovation (Blanz and Ghiselli, 1972) for constructing scales, called the Mixed Standard Scale. As far as we know, neither of these methods has yet been used to evaluate training activities, and the two have not yet been combined into a single system. A combination of the two methods was decided upon as having great potential for our purposes.

The Retranslation Method

The Retranslation Method was developed by Smith and Kendall in an attempt to construct rating statements that are clear, meaningful, and useful to those who actually have to complete the ratings. They observed that most rating procedures are constructed by professionals (usually psychologists) and that the meaning of resulting items usually is not interpreted consistently among those who are expected to make the ratings. They reasoned that statements for rating scales written in the language of the raters themselves

would be less ambiguous and more representative of actual situations than statements developed by psychologists. The important feature of the method, therefore, is the participation of the rater population in deciding what items are important for inclusion in the rating scale and how the rating statements should be written. This method appeared well suited for constructing scales to be used by trainees in assessing Peace Corps training programs, as a common complaint has been that trainees don't always consider the ratings they are asked to make as being very important and there has been confusion among them as to the meaning of scales constructed by the training staff.

The Mixed Standard Scale

The Retranslation Method, as modified for purposes of this study, results only in the production of statements for rating scales. Construction of the scales, or arranging judgments along a quantifiable continuum, is the second part of the problem. The Mixed Standard Scale was first used by Blanz in Finland, and later developed by Blanz and Ghiselli for use in the United States. It was developed in order to minimize the common rating errors of halo and leniency (described previously) as well as to permit an evaluation of the reliability with which each thing is rated, each scale rates, and each rater rates. Reliability ordinarily is determined for the entire process of ratings, which includes the scale, the thing being rated, and the rater. The Mixed Standard scaling procedure makes it possible to differentiate between the accuracy of each, and to inquire into the reliability with which different things are rated, the reliability with which different scales measure, and the reliability of the ratings assigned by different raters.

With most rating procedures the rater is presented with different degrees of "goodness" for each of a number of separate criteria (e.g., trainer skill, clarity of objective) pertaining to a training program, and he selects the one which best describes the program or activity. In the Mixed Standard rating scale there are descriptions of three degrees of each criteria to be rated, and the rater must respond to every description independently. He indicates whether he considers the program or activity to be better than the description, to fit the description, or to be worse than the description.

To reduce the possibility that a rater will form a clear picture of an order of merit set of descriptions for each criterion being rated, the scales and the order of the three statements in them are mixed in a random order. Thus the rater does not deal with all of the statements related to one and the same criterion simultaneously, for he has to rate with respect to each given statement separately. The rater fills out one rating form for each training program or separate activity to be rated. Once the form has been completed, the answers may be rearranged into the form of the commonly employed rating scale so that the questions and answers for each criterion follow one another in order of superiority. By this means, all of the ratings given by the rater on any particular criterion can be viewed simultaneously. By contrast, the rater himself could not do this and could not make a choice among them.

There are two purposes for this procedure. First, the mixing reduces the possibility that the rater will be able to form a clear picture of any order of merit set of descriptions for each criterion being rated. Thus it is anticipated that the errors of halo and leniency will be reduced. Secondly, the mixing provides a means for examining the dependability and reliability

of the ratings, for it permits the ratings to be examined in light of the consistency or logic of the answers to the different statements relating to the same criterion (Blanz and Ghiselli, 1972, pp. 186-187).

The method can be illustrated by means of the following example. Let us assume that the criterion to be rated is "the quality of the training materials used in a particular activity" and has on a scale the following three statements, I being the best description and III the poorest:

I. The materials used in this activity were well prepared and very relevant to the purpose of the activity.

II. The materials used in this activity were adequate and seemed moderately relevant to the purpose of the activity.

III. The materials used in this activity were not well prepared and seemed irrelevant to the purpose of the activity.

If a rater utilizes the procedure accurately, then whenever he checks one statement in a scale as "fits or matches the training activity under consideration" (0), all statements in that scale which describe the activity as inferior will be checked as "the activity was better than the statement" (+). If all three statements are checked with a (+), it means that in the rater's opinion the activity is very good in this criterion, for the activity was superior even to the very best of the three descriptions. Similarly, if all three statements in a scale are checked with a (-), it means that in the rater's opinion the activity was very poor, for the activity was inferior even to the very poorest of the three descriptions.

With the three graded statements used in this manner, there is actually a seven-point scale on each criterion, which also is an improvement on ordinary rating scales. Pursuant to the logic of the system, the various combi-

nations of faultless responses to the items can be arranged as follows and can be assigned the number of points indicated:

Table 1

Descriptive Statement Scale

Descriptive Statements			Points
I	II	III	
+	+	+	7
0	+	+	6
-	+	+	5
-	0	+	4
-	-	+	3
-	-	0	2
-	-	-	1

\+ = The ratee is better than the statement.
0 = The statement fits the ratee.
\- = The ratee is worse than the statement.

The foregoing combinations are faultless because there are no reversals in the order with which the three graded descriptions are checked. That is, whenever a statement is checked with a (0), no statement which describes better performance is checked either (0) or (+) and no statement which describes inferior performance is checked either (0) or (-). Furthermore, (0), which means the statement fits the activity, is not employed for two or more statements which describe degrees of the criterion. All combinations of responses to the three statements, other than the seven given above, are illogical and inconsistent and therefore in error. Nevertheless, the logic of the system permits such scales to be scored. The scoring system for inconsistent responses, that is, where the ratings are in error, are given on the following page.

Table 2

Points to be Assigned for Combinations of Responses Which are Not Logical, and Therefore are in Error

Combination			Points
I	II	III	
+	+	0	7
+	+	-	7
0	+	0	6
0	+	-	6
-	+	0	5
-	+	-	5
0	-	+	5
0	0	-	4
+	0	+	4
+	0	-	4
0	0	0	4
-	0	-	3
+	-	+	3
+	0	0	3
0	-	0	2
+	-	0	2
+	-	-	1
0	-	-	1

+ = Ratee is better than this statement.
0 = Statement fits the ratee.
- = Ratee is worse than this statement.

The determination of the consistency of the ratings, i.e., the number of errors, in fact amounts to a scalogram analysis (Edwards, 1957 and Torgerson, 1958). A variety of different sorts of error counts can be made depending upon the type of accuracy with which one is concerned. Counts can be made of the number of errors per activity, the number of errors per scale, and the number of errors per rater (Blanz and Ghiselli, 1972, pp. 187-188).

The exact nature of these two methods, in terms of our modification and combination of them, is best described according to the steps followed in this project and the scales that resulted from our work. For reasons described later, it was necessary to develop two separate evaluation scales: one for use in assessing the overall training program according to critical criteria defined by trainees, and the other for use in assessing the effectiveness of specific training activities according to criteria defined by professional trainers.

Development of Training Program Evaluation Scales

Step 1. Identification of Training Assessment Criteria

The first step in developing scales for use by trainees in assessing the effectiveness of training according to the Retranslation Method would have been to obtain from trainees statements describing important assessment criteria. Since there were no trainees engaged in training at the time of this study, this was not possible. It was decided, therefore, to sample Volunteers who had participated at different times in several different training programs. Volunteers located in Natal, Rio Grande do Norte; Salvador, Bahia; and Rio de Janeiro were identified for participation in developing the scales. These Volunteers had experienced several different training programs and had been out of training anywhere from one week to one and one-half years.

Nineteen Volunteers were interviewed in Natal for purposes of identifying assessment criteria. Information was obtained in small discussion groups as they responded to the following questions:

- What are the important dimensions of training?
- What were the things that influenced your learning?

- What good things happened in training that helped you adapt to Brazil?
- Can you recall the things that hindered your learning?
- What training experiences seemed most important to you?

The Volunteers were asked to couch their responses in terms of use for evaluation purposes. A large amount of information was thus elicited and documented for later use.

Step 2. Drafting Criterion Statements

The information gathered in Step 1 was organized into a series of statements, each representing a separate idea generated by the Volunteers.

Step 3. Checking the Accuracy of Criterion Statements

The Volunteers who had participated in the original interviews were interviewed a second time. They were asked to examine the draft statements for accuracy in reflecting their views, to make changes where necessary, and to expand the list if important things were left out.

Step 4. Cross-checking the Accuracy of Criterion Statements

In order to generalize and provide a cross-validation for the criterion statements, a second group of Volunteers was interviewed in Salvador and a third group in Rio de Janeiro. They were asked to examine the statements for accuracy in reflecting their views about training, to make necessary modifications, and to expand the list. In addition to these three groups of Volunteers, a number of Peace Corps/Brazil staff members, as well as training staff personnel, were consulted for further checking the accuracy and clarity of the criterion statements.

Step 5. Ranking the Criterion Statements in Order of Importance

A total of twenty-four Volunteers in Natal and Salvador were asked to rank the fifty-eight statements in order of importance. Each Volunteer worked alone and recorded his choices on an individual record card. A list of these statements is presented below in the resulting rank order, along with the rank value. The weighted rank value was computed by adding the total number of ranks given to each statement and dividing this total by the number of Volunteers providing the ranks.

Rank	Rank Value	Statement
(1)	79	Trainers have professional competence in the participative and experiential methods of training.
(2)	73	Training staff is together, well-coordinated, task-oriented toward helping trainee.
(3)	73	Learning climate is psychologically free from rigid interactions and conducive to motivated learning in a flexible setting.
(4)	70	Language teaching method has sufficient variety of techniques to consistently motivate language learning.
(5)	68	Learning climate is physically good and conducive to serious study.
(6)	68	Job information specific and accurate.
(7)	68	Staff is oriented toward helping and supporting each trainee.
(8)	67	Director is easy to talk to, and approach, outside of class.
(9)	67	Training program is primarily experiential, emphasizing, from the first week, learning through experiences outside the center.
(10)	65	Language staff easy to approach and talk to outside of class.
(11)	63	Language and cross-cultural training are coordinated through well chosen training center and off-center experiences.

(12)	62	Criticisms of program easily and pleasantly accepted by language staff.
(13)	61	Flexible pickup on trainee suggestions.
(14)	61	Trainees are offered adequate opportunities to sample off-center cultural activities.
(15)	60	Allocation of trainee time is compatible with trainee energy levels and leisure time needs.
(16)	60	Trainer is a "facilitator," on equal social status with trainee, encouraging trainee to manage his own training plan.
(17)	60	Cross-cultural training emphasizes "openess" and approaches that will serve as guidelines.
(18)	60	The training program appears relevant to the Brazil Peace Corps Volunteer program realities.
(19)	57	Technical orientation has realistic, job-centered, objectives.
(20)	57	Brazilian and American staff both follow the same training philosophy.
(21)	56	The Director participates fully in the training program, making himself easily accessible to trainees.
(22)	56	Supplemental language activities and materials stimulating to trainees, motivating language learning.
(23)	56	Trainees learn "social expectations," etiquette, common mannerisms and behaviors.
(24)	56	Trainee feels he is trusted to learn at his own pace in language training.
(25)	55	Each trainee has a staff member to chat with, someone with whom to share misgivings, doubts, and anger about Peace Corps and training, without fear of defensiveness or reprisal.
(26)	55	Trainees participate in the planning of the weekly activities.
(27)	54	Trainee suggestions immediately discussed and acted upon.
(28)	54	Peace Corps Volunteer job is clearly defined.
(29)	54	Staff includes trainees in planning sessions for the next week.
(30)	53	Defensiveness of key staff members hinders learning, destroys free give and take of the learning climate.

(31)	53	Specialists receive special career consideration in scope of Peace Corps job description.
(32)	53	Trainee helps to firm up his own job description through site visits and Peace Corps programmer help.
(33)	52	Trainee experience in pensao is helpful to the goal of adaptation to Brazil.
(34)	51	Trainee feels he is trusted to learn at his own pace.
(35)	50	Trainee manages his activities toward approaching his site and re-negotiating his job description.
(36)	50	Defensiveness of key staff members discourages suggestions from trainees, limits participation.
(37)	50	Trainees trust the trainers to competently lead them into adaptation to Brazilian life.
(38)	49	Training staff facilitate trainee relationships with Peace Corps Brazil staff by helping to build rapport during training.
(39)	49	Health referrals expedited for those trainees who need professional care.
(40)	48	After one month, trainees visit site and rewrite enroute and terminal objectives.
(41)	48	The estagio experience hastened adaptation to Brazilian life.
(42)	47	Library resources and handouts adequate for learning needs.
(43)	45	The estagio experience improved communication skills with Brazilians.
(44)	44	Trainer follows Peace Corps programmer lead for first site visit and verification of job description.
(45)	44	Trainees assigned to homes during first two weeks.
(46)	43	Trainees help to write their own nucleos for language study.
(47)	43	Language ratings on interim objectives are good learning devices.
(48)	43	Trainees help to write critical incidents on basis of off-center experience.

(49)	40	Trainee's allowance is adequate for his needs.
(50)	40	Women's adaptation needs are addressed specifically.
(51)	38	Critical incidents hit at major issues of adaptation.
(52)	37	Trainees may invite Brazilians to frequent parties -- where trainers also attend.
(53)	33	Terminal objectives are being accomplished in the ten weeks allotted.
(54)	32	Trainee gives effective report of his estagio.
(55)	31	Spouses treated as adults and full Peace Corps members.
(56)	30	Enroute objectives are being met on schedule.
(57)	29	The training design is being implemented according to its original plan.
(58)	27	Trainee learns to write a business letter, use banking forms.

This rank distribution was then examined, and the top twenty items were selected for further development (ranked items 1 through 20).

Step 6. Construction of Criterion Statement Categories

The twenty statements selected in Step 5 were studied to determine the dimensions of training assessment criteria they represented. This resulted in the construction of the following fourteen categories:

1. Training staff expertise in applying Peace Corps methodology
2. Training staff team performance
3. Training staff availability to trainees
4. Training program director availability and responsiveness
5. Experiential learning based on host community environment
6. Training staff responsiveness to trainee suggestions
7. Cross-cultural training method
8. Language training method
9. Coordination of resources of individual needs
10. Realistic job-centered objectives

11. Accurate job descriptions
12. Opportunities to sample off-training-center Brazilian cultural activities
13. Physically adequate learning climate
14. Training schedule

It was decided that Categories 1-9 (labeled A through I in our scales) would be developed into rating scales according to the Mixed Standard scaling method, while 10-14 would be included as "yes" or "no" questions.

Step 7. Drafting Degrees of Effectiveness for Criterion Categories

Each of the first nine categories constructed in Step 6 was studied separately for the purpose of developing rating items. A set of three statements was written for each category, the first representing high effectiveness, the second representing medium effectiveness, and the third representing low effectiveness.

Step 8. Checking for Accurate Inclusion of Effectiveness Statements in Original Categories

A group of eleven judges was selected to determine whether the sets of effectiveness statements were perceived as clearly belonging to the original categories (an important consideration in determining scale errors in the Mixed Standard scaling procedure). The twenty-seven effectiveness statements were arranged in random order on cards, and the nine categories were written on cards and placed side by side. The judges were asked to place each statement in the category they thought it clearly represented. The results of the judging indicated a high accuracy in the statements being placed in their original categories. There were some errors made in Categories A, C, F, and I, however, so the ambiguous statements causing these errors were modified.

Step 9. Checking for Accurate Degrees of Effectiveness for Criterion Categories

The same eleven judges who checked for accurate inclusion of effectiveness statements in original categories were used to check the accuracy of the draft statements indicating degrees of effectiveness for each of the categories. This was done by arranging the twenty-seven statements in random order and having the judges place each set of three statements in their original categories as indicated in Step 8. When each set of three statements was placed in its proper category, the judges were asked to order the three statements from high to low according to their perception of the degree of effectiveness represented. Again the results of this judging indicated a high degree of accuracy in the draft statements being written in the proper order of high, medium, and low effectiveness. There were some errors made in Categories A, C, F, and I, so the ambiguous statements causing these errors were rewritten.

The twenty-seven statements arranged in their proper order of effectiveness and in the proper categories, resulting from the procedures described in Steps 1-9, are listed below:

Category A. Training Staff Expertise in Applying Peace Corps Training Methodology

I. The majority of training staff competently apply appropriate Peace Corps training methodology, showing kindness and consistency in the way they deal with trainees.

II. About half of the training staff competently apply appropriate Peace Corps training methodology, showing kindness and consistency in the way they deal with trainees.

III. Only a few trainers competently apply appropriate Peace Corps training methodology, showing kindess and consistency in the way they deal with trainees.

Category B. Training Staff Team Performance

I. The staff appear to have a good team approach toward conflict resolution and building a favorable learning climate of open interaction among themselves and with trainees.

II. The staff appear to have a divided team approach, some inability in resolving conflict, and fair success in building a learning climate of open interaction among themselves and with trainees.

III. The staff appear to lack a team approach, affect the whole center with their conflict, and/or segment the learning climate according to the philosophy of each trainer.

Category C. Training Staff Availability to Trainees

I. Most training staff seek extra time opportunities for talking with individual trainees.

II. Most training staff spend extra time talking with individual trainees.

III. Most training staff avoid spending extra time with individual trainees.

Category D. Training Program Director Availability and Responsiveness

I. The Director is easily approachable, and often participates in informal group and individual discussions with trainees.

II. The Director is sort of approachable, and occasionally participates in informal group and individual discussions with trainees.

III. The Director is difficult to approach and rarely participates in informal group or individual discussions with trainees.

Category E. Experiential Learning Based on Host Community Environment

I. Guided learning activities in the community are scheduled as often as possible, and are well integrated into the total learning program.

II. Learning activities in the community are occasionally scheduled; these activities are usually integrated into the total learning program, but sometimes suffer from lack of staff guidance.

III. Learning activities in the community are rarely scheduled. These activities suffer from lack of staff guidance and are poorly integrated into the training program.

Category F. Training Staff Responsiveness to Trainee Suggestions

I. Training staff seek feedback from trainees and always deal with suggestions and criticisms immediately to mutually find the best solution.

II. Training staff seek feedback from trainees and usually deal with suggestions and criticisms, but seldom take immediate action to work out changes or solutions.

III. Training staff avoid feedback from trainees and rarely deal with suggestions and criticisms in such a way as to make changes or find solutions.

Category G. Cross-cultural Training Method

I. Cross-cultural training emphasizes a variety of alternative behaviors that are appropriate to specific situations, utilizing the larger theoretical and cultural context for greater understanding.

II. Cross-cultural training includes some variety of alternative behaviors that are appropriate to specific situations, and utilizes the historical context for greater understanding.

III. Cross-cultural training is restricted to prescriptions of stereotyped behavior that are appropriate or inappropriate to the Brazilian culture.

Category H. Language Training Method

I. Language trainers use a great variety of teaching techniques that consistently contribute to individual motivation for language learning.

II. Language trainers use some variety in teaching techniques, but not sufficiently to consistently contribute to individual motivation for language learning.

III. Language trainers use little variety in teaching techniques, which contributes to loss of motivation for language learning.

Category I. Coordination of Resources for Individual Needs

I. Most training staff consistently coordinate their training activities and resources to address the needs of an individual trainee.

II. Some training staff coordinate their training activities and resources to address the needs of an individual trainee.

III. Most training staff tend to view trainees as a group and rarely coordinate their training activities and resources to address the needs of an individual trainee.

Step 10. Application of the Mixed Standard Scaling Method

The final step in the procedure was to apply the Mixed Standard Scaling Method to the categories and statements resulting from previous steps. The statements were arranged in random order and incorporated into the rating scale, with proper instructions, scoring procedures, and methods for checking reliability, all of which is presented in Chapter III of this report.

Development of Training Activity Evaluation Scales

As stated earlier, this project was initiated in order to develop scales for use in evaluating the effectiveness of training activities. The procedures described in the ten steps above were followed to achieve this objective. However, an examination of the nine criterion categories and the twenty-seven effectiveness statements that resulted from this procedure clearly indicates that they are not suitable for purposes of assessing specific training activities. These criteria are much broader in nature, and have to do with the effectiveness of the training experience as a whole. They represent what trainees think are the most important aspects of Peace Corps training in general, not elements of successful training activities, exercises, or lessons. (It was significant, and somewhat surprising, to discover the overwhelming importance trainees placed on the training staff. Six of the fourteen criterion categories were in direct reference to the training staff.)

It was decided, therefore, that these items would be retained for use in evaluating training programs in general -- perhaps at the middle and again at the conclusion of training -- and that a second set of criteria would be developed to assess training activities. It was also decided that professional trainers, rather than trainees, would be used to develop these statements as trainers have a better understanding of the technical elements that combine to characterize a successful training activity.

The five steps involved in this procedure are outlined below:

Step 1. Identification of Training Activity Assessment Criteria

Five professional Peace Corps trainers were brought together to identify criteria of effective training activities. They were asked to brainstorm all the elements of any given training activity (e.g., language class, case study exercises, field experience) that contributed to its effectiveness or ineffectiveness. Each trainer was then asked to write down the five elements he thought were most important. A final set of five criteria was selected by examining the combined lists of the five trainers. These five criteria were:

A. Clarity of the objective

B. Skill of the trainer

C. Effectiveness of the method

D. Quality of materials

E. Subjective estimate of learning achieved related to Volunteer service

Step 2. Drafting Degrees of Effectiveness for Criterion Categories

Each trainer was then asked to write a set of three effectiveness statements for each of the five criterion categories resulting from Step 1.

The set was to represent three degrees of effectiveness for each category -- high, medium, and low. These statements were used to produce a final set of fifteen statements (a set of three for each of the five criterion categories).

Step 3. Checking for Accurate Inclusion of Effectiveness Statements in Criterion Categories

The same eleven judges who participated in developing the Training Program Evaluation Statements were employed to determine whether the sets of effectiveness statements were clearly perceived as belonging to the criterion categories for which they were written. The fifteen effectiveness statements were arranged in random order on cards, and the five categories were written on cards and placed side by side. The judges were asked to place each effectiveness statement in the category they thought it clearly represented. The results of this judging indicated a high accuracy in the statements being placed in their original categories. One of the eleven judges indicated some difficulty distinguishing between Category B and Category C. He felt that the method can only be as good as the trainer. However, most of the judges had no difficulty; the items seemed to fit easily into the categories.

Step 4. Checking for Accurate Degrees of Effectiveness Statements in Criterion Categories

The eleven judges were then asked to participate in checking the accuracy with which the fifteen statements represented high, medium, and low degrees of effectiveness in each category. After the statements had been placed in their proper category (in Step 3), the judges were asked to arrange them in order of effectiveness represented from high to medium to low. The results of this judging indicated a high degree of accuracy. Ten

of the eleven judges placed the statements correctly in the effectiveness sequence for which they were written.

The fifteen statements resulting from this procedure, in the proper categories and order of merit sequence, are listed below.

Category A. Clarity of Objective

I. I clearly understood the objective of this activity.

II. I think I understood the objective of this activity, but it's not too clear.

III. I did not understand the objective of this activity.

Category B.. Skill of the Trainer

I. The trainer was very skillful in conducting this activity. The effective use of these skills greatly facilitated my learning.

II. The trainer conducted this activity fairly well, but could have used more skill in helping me learn.

III. The trainer was not skillful in conducting this activity, and did not help me learn.

Category C. Effectiveness of Method

I. The method used in implementing this activity was very effective in facilitating my learning.

II. The method used in implementing this activity was all right, but it did not particularly facilitate my learning.

III. The method used in implementing this activity did not facilitate my learning.

Category D. Quality of Materials

I. The materials used in this activity were well prepared and very relevant to the purpose of this activity.

II. The materials used in this activity were adequate and seemed moderately relevant to the purpose of the activity.

III. The materials used in this activity were not well prepared and seemed irrelevant to the purpose of the activity.

Category F. Subjective Estimate of Learning Achieved Related to Volunteer Service

I. I learned a great deal from this activity which I feel has helped me prepare for Volunteer service.

II. I learned a moderate amount from this activity and some of what I learned has helped me prepare for Volunteer service.

III. I learned very little from this activity, and I don't think it has helped me prepare for Volunteer service.

Step 5. Application of the Mixed Standard Scaling Method

The final step in the procedure was to apply the Mixed Standard Scaling Method to the categories and statements resulting from the first four steps. The fifteen statements were arranged in random order and incorporated into the rating scale with proper instructions, scoring procedures, and methods for checking reliability, all of which is presented in Chapter III of this report.

CHAPTER III. PRODUCTS

The Training Program Evaluation Scales

The complete system for evaluating training programs appears at the end of this section, beinning on Page 29. The scales themselves should be administered exactly as they appear and as described in the instructions. It is recommended that these scales be used once or twice during the length of a training program -- about midway through the program and again near the program's conclusion. It is important that the resulting evaluation information be immediately shared among the training community for program modification and improvement, as well as the data being systematically analyzed and stored in a central location (Peace Corps/Brazil or Peace Corps/Washington).

Scoring

(A look at the evaluation scales, beginning on Page 29, at this point will probably make the following description easier to understand.)

The last five items in the scales are straightforward and simple to score. The total number of "yes," "a little" and "no" responses are simply examined and tabulated for each of the five items.

The other twenty-seven statements are more complicated and require a somewhat elaborate scoring procedure. These statements were arranged in random order so that the rater could not easily determine which statements belonged to a particular category or which statement fit in an order-of-effectiveness sequence. These statements must be rearranged in the proper sequence and category for scoring purposes. This is done by examining the Scoring Keys on Page 34 and the Scoring Matrix on Page 36. The items belonging to each category, according to the Key, are examined on each response sheet, and the scale value is found on the Scoring Matrix.

For example, for Criterion Category A - Training Staff Expertise in Applying Peace Corps Training Methodology - items #9, #27, and #18 are examined in that sequence. If a given response set to these three items is #9 = +, #27 = +, #18 = +, the resulting score is 7. If another response set to these items is #9 = -, #27 = -, #18 = +, the resulting score is 3, and so on for all possible response combinations as indicated in the Scoring Matrix.

Each resulting scale value should be listed for all respondents on the Scoring Work Sheet on Page 37. When all of the nine scales have been scored for all of the respondents, the scale sums and scale means should be computed as indicated at the bottom of the Work Sheet. Each of these nine scale means should then be listed on the Summary Sheet on Page 38, along with the frequency tabulations on the last five questions and a summary of comments and suggestions. Copies of the Summary Sheet should be distributed to all staff and trainees for feedback and discussion. This Summary Sheet should also be used for purposes of program evaluation record-keeping.

Error Computation

One of the strong assets of the Mixed Standard Scaling Method is that it is possible to keep an immediate and continuous accounting of the accuracy of the rating procedure. Each time the scales are used, it is possible to compute the rater errors as well as the individual scale errors. This is done by using the Error Computation Work Sheet on Page 39. The responses on each of the nine scales are listed, using the Scoring Keys and the Scoring Matrix. Then, the errors are listed for each set of ratings in each scale by referring to the Error Key appearing on Page 42. Any set of ratings other than the seven appearing in the Error Key are in error according to the

logic of the system. Each error is noted and listed in the appropriate space on the Work Sheet. The errors for each rater are computed by summing the errors made by each rater across the nine scales. The average rater error is computed by summing each rater's error and dividing by the number of raters. A rater error percentage can be computed by dividing the average rater error by 9.

The errors made in using each of the nine scales can be computed by summing the total number of errors made by the respondents on each scale. A scale error percentage can be computed by dividing the sum of each scale error by the number of respondents.

A complete example of how the error computation system works is presented on Page 43. There are several ways that this information can be interpreted and used. Referring to the example, these are:

- Scale 3 was the most reliable, since there were no errors made in using this scale. On the other hand, Scale 1 was the most unreliable, showing a 50% error rate (half of the respondents made an error in using this scale). The other seven scales fall between these two extremes. In general, the smaller the error, the more confidence there is for decision making in using the results of ratings made on the scale.
- Raters #10 and #20 were the most reliable, since they made no errors in using all nine scales. Raters #5, #8, and #13 were the most unreliable, since they made errors in four out of the nine scales. The scale data could be made more reliable by not using the ratings made by the most unreliable raters.
- The overall rater error was 26%, or conversely, the overall rater accuracy was 74%. This figure provides a general idea of the accuracy of the evaluating system. No method has yet been developed (of which we are aware) whereby this figure could be

translated to a reliability correlation coefficient, so it cannot be interpreted as a reliable coefficient normally would. However, it does provide a valid and quick estimate of the accuracy of rating data, which should be carefully considered prior to decision making based on rating outcomes.

Training Program Evaluation Scales

Instructions

The scales on the following pages were constructed in order to assess the effectiveness of several critical aspects of the training program. Please complete the scales exactly as instructed below:

1. Complete the scales by responding to each Descriptive Statement, one by one, in the order in which they are presented (1-27).
2. The ratings are to be made in the following manner: Consider each Descriptive Statement independently from the others and decide whether you think the training program being evaluated was worse than the Descriptive Statement; matched the statement; or was better than the statement.
   -- If you think the program was worse than the statement, place a - mark in the corresponding box.
   -- If you think the program matched the statement, place a 0 mark in the corresponding box.
   -- If you think the program was better than the statement, place a + mark in the appropriate box.
3. When you have finished responding to all 27 statements in the above manner, respond to the additional 5 statements listed on the last page by placing a check (✓) mark in either the "yes," "a little," or "no" box.
4. Write any comments or suggestions you may have on the back of the paper.

TRAINING PROGRAM EVALUATION SCALES

Name _____ Date _____

Descriptive Statements

		Responses (-, 0, +)
1.	Most training staff tend to view trainees as a group and rarely coordinate their training activities and resources to address the needs of an individual trainee.	
2.	Language trainers use some variety in teaching techniques but not sufficiently to consistently contribute to individual motivation for language learning.	
3.	Cross-cultural training emphasizes a variety of alternative behaviors that are appropriate to specific situations, utilizing the larger theoretical and cultural context for greater understanding.	
4.	Training staff avoid feedback from trainees and rarely deal with suggestions and criticisms in such a way as to make changes or find solutions.	
5.	Learning activities in the community are occasionally scheduled; these activities are usually integrated into the total learning program, but sometimes suffer from lack of staff guidance.	
6.	The Director is easily approachable, and often participates in informal group and individual discussions with trainees.	
7.	Most training staff avoid spending extra time with individual trainees.	
8.	The staff appear to have a divided team approach, some inability in resolving conflict, and fair success in building a learning climate of open interaction among themselves and with trainees.	
9.	The majority of training staff competently apply appropriate Peace Corps training methodology, showing kindness and consistency in the way they deal with trainees.	
10.	Some training staff coordinate their training activities and resources to address the needs of an individual trainee.	
11.	Language trainers use a great variety of teaching techniques that consistently contribute to individual motivation for language learning.	

[Write any comments or suggestions you may have on the back of this paper.]

Descriptive Statements	Responses (-, 0, +)	
12. Cross-cultural training is restricted to prescriptions of stereotyped behavior that are appropriate or inappropriate to the Brazilian culture.		[Write any comments or suggestions you may have on the back of this paper.]
13. Training staff seek feedback from trainees and usually deal with suggestions and criticisms, but seldom take immediate action to work out changes or solutions.		
14. Guided learning activities in the community are scheduled as often as possible, and are well integrated into the total learning program.		
15. The Director is difficult to approach and rarely participates in informal group or individual discussions with trainees.		
16. Most training staff spend extra time talking with individual trainees, but only when trainees approach staff members.		
17. The staff appears to have a good team approach toward conflict resolution and building a favorable learning climate of open interaction among themselves and with trainees.		
18. Only a few trainers competently apply appropriate Peace Corps training methodology, showing kindness and consistency in the way they deal with trainees.		
19. Most training staff consistently coordinate their training activities and resources to address the needs of an individual trainee.		
20. Language trainers use little variety in teaching techniques, which contributes to loss of motivation for language learning.		
21. Cross-cultural training includes some variety of alternative behaviors that are appropriate to specific situations, and utilizes the historical context for greater understanding.		
22. Training staff seek feedback from trainees and always deal with suggestions and criticisms immediately to mutually find the best solution.		
23. Learning activities in the community are rarely scheduled. These activities suffer from lack of staff guidance and are poorly integrated into the training program.		

Descriptive Statements

Responses (-, 0, +)

24. The Director is sort of approachable, and occasionally participates in informal group and individual discussions with trainees.

25. Most training staff seek extra time opportunities for talking with individual trainees.

26. The staff appear to lack a team approach, affect the whole center with their conflict, and/or segment the learning climate according to the philosophy of each trainer.

27. About half of the training staff competently apply appropriate Peace Corps training methodology, showing kindness and consistency in the way they deal with trainees.

[Write any comments or suggestions you may have on the back of this paper.]

Part II

Yes	A Little	No	
☐	☐	☐	The training program has realistic job-centered objectives relevant to the Peace Corps Volunteer program in Brazil.
☐	☐	☐	Job descriptions contain specific information and are up-dated for accuracy.
☐	☐	☐	Trainees are offered adequate opportunities to sample off-training-center cultural activities during leisure time.
☐	☐	☐	The learning climate is physically good, conducive to serious study.
☐	☐	☐	The training schedule is compatible with trainee energy levels and leisure time needs.

TRAINING PROGRAM EVALUATION SCALES

Scoring Keys

Scoring Key 1. Training Staff Expertise in Applying Peace Corps Training Methodology

(High) I = Item 9
(Medium) II = Item 27
(Low) III = Item 18

Scoring Key 2. Training Staff Team Performance

I = Item 17
II = Item 8
III = Item 26

Scoring Key 3. Training Staff Availability to Trainees

I = Item 25
II = Item 16
III = Item 7

Scoring Key 4. Training Program Director Availability and Responsiveness

I = Item 6
II = Item 24
III = Item 15

Scoring Key 5. Experiential Learning Based on Host Community Environments

I = Item 14
II = Item 5
III = Item 23

Scoring Key 6. Training Staff Responsiveness to Trainee Suggestions

I = Item 22
II = Item 13
III = Item 4

Scoring Key 7. Cross-Cultural Training Method

I = Item 3
II = Item 21
III = Item 12

Scoring Key 8. Language Training Method

I = Item 11
II = Item 2
III = Item 20

Scoring Key 9. Coordination of Resources for Individual Needs

I = Item 19
II = Item 10
III = Item 1

TRAINING PROGRAM EVALUATION SCALES

Scoring Matrix

Descriptive Statements			Points
I (High)	II (Medium)	III (Low)	
+	+	+	7
+	+	0	7
+	+	-	7
0	+	+	6
0	+	0	6
0	+	-	6
-	+	+	5
-	+	0	5
-	+	-	5
0	-	+	5
-	0	+	4
0	0	-	4
+	0	+	4
+	0	-	4
0	0	0	4
-	-	+	3
-	0	-	3
+	-	+	3
+	0	0	3
-	-	0	2
0	-	0	2
+	-	0	2
-	-	-	1
+	-	-	1
0	-	-	1

+ = The program was <u>better than</u> the statement.
0 = The statement <u>matches</u> the program.
- = The program was <u>worse than</u> the statement.

TRAINING PROGRAM EVALUATION SCALES

Scoring Work Sheet

Scales →	1. Training Staff Expertise	2. Staff Team Performance	3. Staff Availability	4. Director Responsiveness	5. Experiential Learning	6. Staff Responsiveness	7. Cross-Cultural Training	8. Language Training	9. Coordination for Individual Needs
Trainees									
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
Sums	1.	2.	3.	4.	5.	6.	7.	8.	9.
Means $\left(\frac{\text{Sum}}{N}\right)$	1.	2.	3.	4.	5.	6.	7.	8.	9.

Write these values on the summary sheet.

TRAINING PROGRAM EVALUATION SCALES

Summary Sheet

Training Program _____

Date _____

Scale 1. Training Staff Expertise in Applying Peace Corps Training Methodology — Mean Score _____

Scale 2. Training Staff Team Performance — Mean Score _____

Scale 3. Training Staff Availability to Trainees — Mean Score _____

Scale 4. Training Program Director Availability and Responsiveness — Mean Score _____

Scale 5. Experiential Learning Based on Host Community Environments — Mean Score _____

Scale 6. Training Staff Responsiveness to Trainee Suggestions — Mean Score _____

Scale 7. Cross-Cultural Training Method — Mean Score _____

Scale 8. Language Training Method — Mean Score _____

Scale 9. Coordinator of Resources for Individual Needs — Mean Score _____

	Yes	A Little	No
Realistic, Job-Centered Objectives			
Job Descriptions			
Brazilian Cultural Activities			
Learning Climate			
Training Schedule			

Summary of Comments and Suggestions

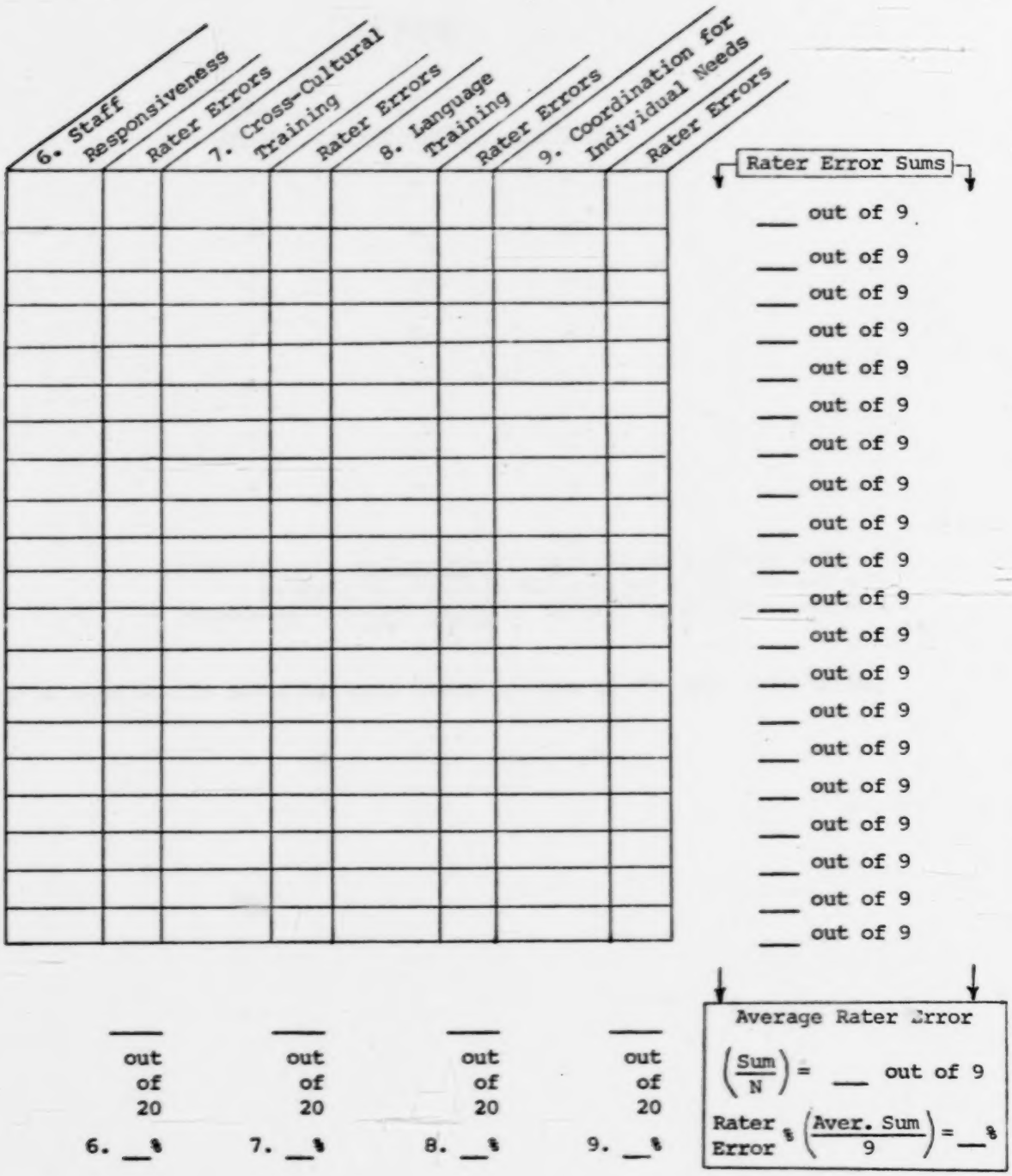
TRAINING PROGRAM EVALUATION SCALES

Error Computation Work Sheet

Trainees	1. Training Staff Expertise	Rater Errors	2. Staff Team Performance	Rater Errors	3. Staff Availability	Rater Errors	4. Director Responsiveness	Rater Errors	5. Experiential Learning	Rater Errors
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										

Scale Error Sums = ____ out of 20 ____ out of 20 ____ out of 20 ____ out of 20 ____ out of 20

Scale Error % $\left(\frac{\text{Sum}}{20}\right)$ = 1. __% 2. __% 3. __% 4. __% 5. __%


6. Staff Responsiveness
Rater Errors
7. Cross-Cultural Training
Rater Errors
8. Language Training
Rater Errors
9. Coordination for Individual Needs
Rater Errors
Rater Error Sums
___ out of 9
___ out of 9
___ out of 9

ERROR KEY

I (High)	II (Medium)	III (Low)
+	+	+
0	+	+
-	+	+
-	0	+
-	-	+
-	-	0
-	-	-

+ = The activity or program is <u>better than</u> the statement.
0 = The statement <u>matches</u> the activity or program.
- = The activity or program is <u>worse than</u> the statement.

USE OF KEY: Any set of ratings other than the seven listed here is in error.

TRAINING PROGRAM EVALUATION SCALES

Example of Error Computation

Trainees (Scales)	1. Training Staff Expertise	Rater Errors	2. Staff Team Performance	Rater Errors	3. Staff Availability	Rater Errors	4. Director Responsiveness	Rater Errors	5. Experiential Learning	Rater Errors
1	+++		00-	1	0++		0+-	1	-++	
2	-++		+0+	1	-++		0++		0-+	1
3	0++		+-0	1	-0+		-++		-0+	
4	0+0	1	-0+		0++		0+0	1	-0+	
5	++-	1	-0+		-++		0+0	1	-++	
6	0++		-++		-0+		0++		0-+	1
7	-+0	1	-0+		0++		-++		-0+	
8	-+0	1	--+		-++		0+0	1	-0+	
9	++-	1	0++		-0+		0++		0-+	1
10	+++		--+		0++		-++		-++	
11	+++		000	1	-++		0+0	1	00-	1
12	-++		-0-	1	-0+		0++		-++	
13	0++		+0+	1	0++		0+0	1	--+	
14	0+0	1	0++		-++		-++		--+	
15	++-	1	--+		-0+		0++		--+	
16	0++		000	1	0++		-++		00-	1
17	-+0	1	-0+		-++		-++		-0+	
18	-+0	1	-0+		-0+		-++		-0+	
19	++-	1	-0+		0++		0+-	1	--+	
20	+++		-0+		-++		-++		--+	

Scale Error Sums = 10 out of 20 | 7 out of 20 | 0 out of 20 | 7 out of 20 | 5 out of 20

Scale Error % $\left(\frac{\text{Sum}}{20}\right)$ = 1. 50 % | 2. 35 % | 3. 0 % | 4. 35 % | 5. 25 %

6. Staff Responsiveness	Rater Errors	7. Cross-Cultural Training	Rater Errors	8. Language Training	Rater Errors	9. Coordination for Individual Needs	Rater Errors	Rater Error Sums
0++		-+-	1	-0+		-0+		3 out of 9
0++		-0+		-+-	1	-++		3 out of 9
00-	1	-0+		-0+		-++		2 out of 9
--+		-++		-0+		++0	1	3 out of 9
0-+	1	-++		--+		++-	1	4 out of 9
--+		0++		-+-	1	0++		2 out of 9
-++		-+-	1	--+		0++		2 out of 9
0-+	1	0++		--+		++0	1	4 out of 9
-++		-++		-+-	1	0++		3 out of 9
-0+		0++		--+		+++		0 out of 9
-0+		0++		-++		+++		3 out of 9
0-+	1	0++		-++		+++		2 out of 9
-0+		-+-	1	-++		++-	1	4 out of 9
--+		-++		-0+		0++		1 out of 9
--+		-++		-0+		0++		1 out of 9
-++		0++		-+-	1	0++		3 out of 9
-++		-++		-0+		++0	1	2 out of 9
--+		+-+	1	-0+		+++		2 out of 9
--+		--+		-0+		++-	1	3 out of 9
-++		--+		-0+		+++		0 out of 9

4 out of 20	4 out of 20	4 out of 20	6 out of 20
6. 20%	7. 20%	8. 20%	9. 30%

Average Rater Error

$\left(\frac{\text{Sum}}{N}\right)$ = 2.35 out of 9

Rater Error % $\left(\frac{\text{Aver. Sum}}{9}\right)$ = 26%

The Training Activity Evaluation Scales

The complete system for evaluating specific training activities appears at the end of this section, beginning on Page 49. The scales themselves should be administered exactly as they appear and as described in the instructions. It is recommended that these scales be used at the end of the second week of training, and either each week or every other week thereafter through the conclusion of the program. The four or five major training activities conducted during the evaluation period (e.g., morning language classes, case study exercise, lecture on history) should be listed on the scales before being distributed for completion of ratings. It is important that the resulting evaluation information be immediately shared among the training community for program modification and improvement, as well as the data being systematically analyzed and stored in a central location (Peace Corps/Brazil or Peace Corps/Washington).

Scoring

The fifteen rating statements are arranged in random order on the scales so that the rater cannot easily determine which statements belong to particular categories or which statements fit in an order of effectiveness sequence. The statements must be rearranged in their proper sequence and category for scoring purposes. This is done by examining the Scoring Keys beginning on Page 52 and the Scoring Matrix on Page 57. The items belonging to each category according to the Key are examined on each response sheet and the scale value is found on the Scoring Matrix. Since the scales are designed in such a way that five separate training activities can be evaluated on one page, scoring overlays have been prepared for use in scoring each of the criterion categories as they are presented here. Each of the

five criterion categories can therefore be scored by placing the scoring key overlay for a particular category on a response sheet and noting the scores for that category across each of the activities rated. (The sample keys shown here can be overlaid on each of the two sheets used for the scales.)

For example, for Criterion Category 1 - Clarity of Objective - items #10, #5 and #15 appear in the scoring overlay and are examined in that sequence. If a given response set to these items for Activity A is #10 = +, #5 = +, and #15 = +, the resulting score is 7. If another response set to these items, on Activity B, is #10 = -, #5 = -, and #15 = +, the resulting score is 3, and so on for all possible response combinations as indicated in the Scoring Matrix for each overlay across activities being rated.

Each resulting scale value should be listed for all respondents on the Scoring Work Sheets beginning on Page 58. When all of the five scales have been scored for all of the respondents across all activities being evaluated, the scale sums and scale means for each activity should be computed as indicated on the bottom of the Work Sheets. Each of these five scale means for each of the activities being evaluated should then be listed on the Summary Sheet on Page 63. Copies of the Summary Sheet should be distributed to all staff and trainees for feedback and discussion. This Summary Sheet should also be used for purposes of program evaluation record-keeping.

As described in the previous section for the Training Program Evaluation Scales, each time these Activity Scales are used, it is possible to compute the rater errors as well as the individual scale errors. In this case, however, errors can also be computed for each of the activities being evaluated at any given time. The error computation is done by using the Error Computation Work Sheet on Page 64. Only one Work Sheet has been prepared,

but additional sheets exactly like this one should be prepared for each activity that has been evaluated. The responses on each of the five scales are listed using the Scoring Keys and the Scoring Matrix. Then the errors are listed for each set of ratings and each scale by referring to the Error Key presented on Page 65. Any set of ratings other than the seven appearing in the Error Key are in error according to the logic of the system. Each error is noted and listed in the appropriate space on the Work Sheet. The errors for each rater are computed by summing the errors made by each rater across the five scales. The average rater error is computed by summing each rater error and dividing by the number of raters. A rater error percentage can be computed for each activity being evaluated by dividing the average rater error by five.

The errors made in using each of the five scales can be computed by summing the total number of errors made by the respondents on each scale. A scale error percentage can be computed by dividing the sum of each scale error by the number of respondents. An example of how this error computation system works for these Training Activity Evaluation Scales is not presented. However, the example for the Training Program Evaluation Scales appearing on Page 43 is sufficient for these purposes. Referring to that example, the error information for these scales can be interpreted and used in the same ways described on Page 27. That is, each of the five scale reliabilities can be determined. The errors made by each rater can be determined, and the scale data made more reliable by not using the ratings made by the most unreliable raters. The overall rater error, or conversely, the overall rater accuracy, can be determined for each of the training activities evaluated, which provides a general idea of the accuracy of the evaluation system.

Training Activity Evaluation Scales

Instructions

The scales on the following page were constructed in order to assess the effectiveness of training activities. Please complete the scales exactly as instructed below:

1. Rate each activity one at a time by responding to all fifteen descriptive statements for the first activity, then for the second, third, fourth, and fifth.
2. The ratings are to be made in the following manner: Consider each Descriptive Statement independently of the others, one at a time, in the order listed (1-15). Decide whether you think the training activity being evaluated was worse than the Descriptive Statement; the activity matched the statement; or the activity was better than the statement.
   -- If you think the activity was worse than the statement, place a - mark in the appropriate box.
   -- If you think the activity matched the statement, place a 0 mark in the appropriate box.
   -- If you think the activity was better than the statement, place a + mark in the appropriate box.
3. Write any comments or suggestions you may have on the back of the paper.

TRAINING ACTIVITY EVALUATION SCALES

Name _____

Date _____

Descriptive Statements	Activities A	B	C	D	E
1. I learned very little from this activity and I don't think it has helped me prepare for Volunteer service.					
2. The materials used in this activity were adequate and seemed moderately relevant to the purpose of the activity.					
3. The method used in implementing this activity was very effective in facilitating my learning.					
4. The trainer was not skillful in conducting this activity, and did not help me learn.					
5. I think I understood the objective of this activity, but it was not very clear.					
6. I learned a great deal from this activity which I feel has helped me prepare for Volunteer service.					
7. The materials used in this activity were not well prepared and seemed irrelevant to the purpose of the activity.					
8. The method used in implementing this activity was all right, but it did not particularly facilitate my learning.					
9. The trainer conducted this activity fairly well, but could have used more skill in helping me learn.					
10. I clearly understood the objective of this activity.					
11. I learned a moderate amount from this activity and some of what I learned has helped me prepare for Volunteer service.					
12. The materials used in this activity were well prepared and very relevant to the purpose of this activity.					

[Write any comments or suggestions you may have on the back of this paper.]

Descriptive Statements	Activities A	B	C	D	E
13. The method used in implementing this activity did not facilitate my learning.					
14. The trainer was very skillful in conducting this activity. The effective use of these skills greatly facilitated my learning.					
15. I did not understand the objective of this activity.					

[Write any comments or suggestions you may have on the back of this paper.]

TRAINING ACTIVITY EVALUATION SCALES

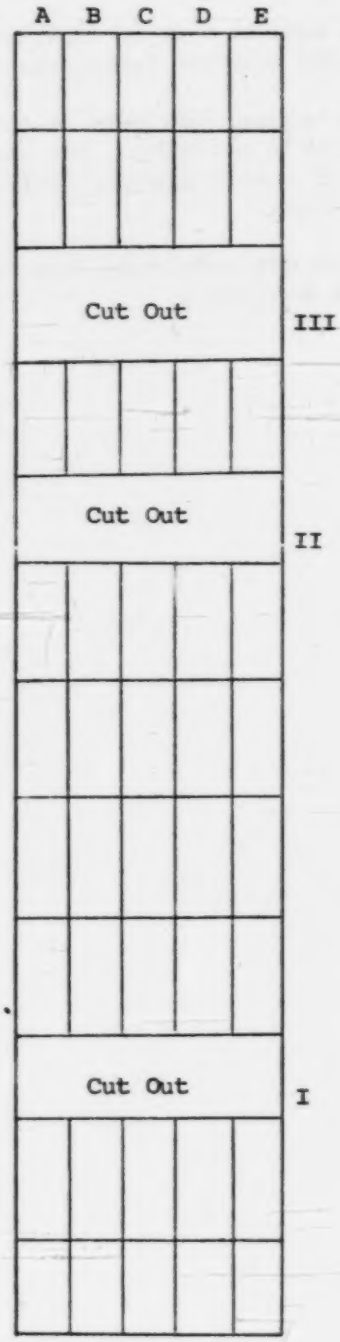
Scoring Keys

Scoring Key 1. Clarity of Objective

15. I did not understand the objective of this activity.

5. I think I understood the objective of this activity, but it's not too clear.

10. I clearly understood the objective of this activity.

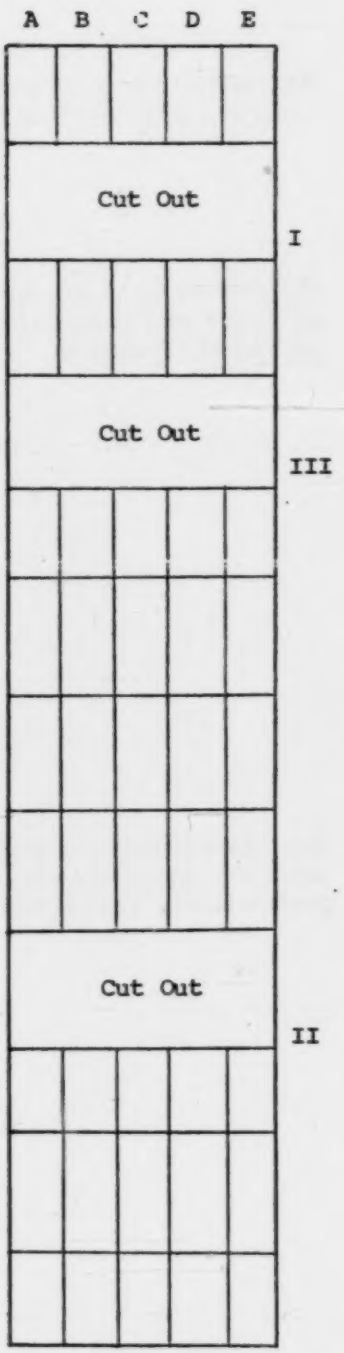


Scoring Key 2. Skill of the Trainer

14. The trainer was very skillful in conducting this activity. The effective use of these skills greatly facilitated my learning.

4. The trainer was not skillful in conducting this activity, and did not help me learn.

9. The trainer conducted this activity farily well, but could have used more skill in helping me learn.

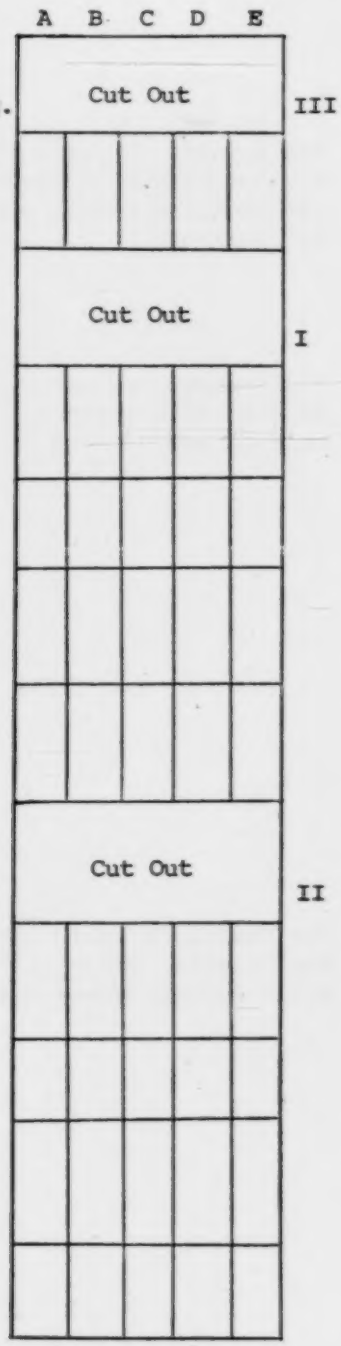


Scoring Key 3. Effectiveness of Method

13. The method used in implementing this activity did not facilitate my learning.

3. The method used in implementing this activity was very effective in facilitating my learning.

8. The method used in implementing this activity was all right, but it did not particularly facilitate my learning.

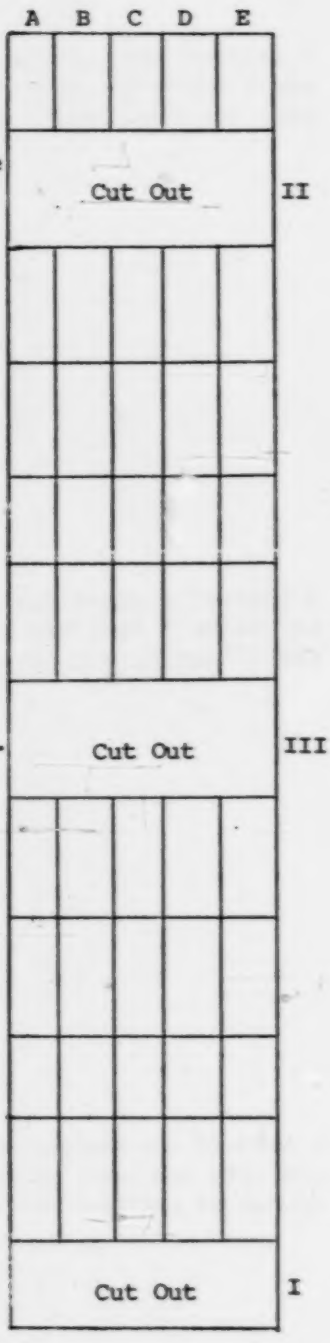


Scoring Key 4. Quality of Materials

2. The materials used in this activity were adequate and seemed moderately relevant to the purpose of the activity.

7. The materials used in this activity were not well prepared and seemed irrelevant to the purpose of the activity.

12. The materials used in this activity were well prepared and very relevant to the purpose of this activity.

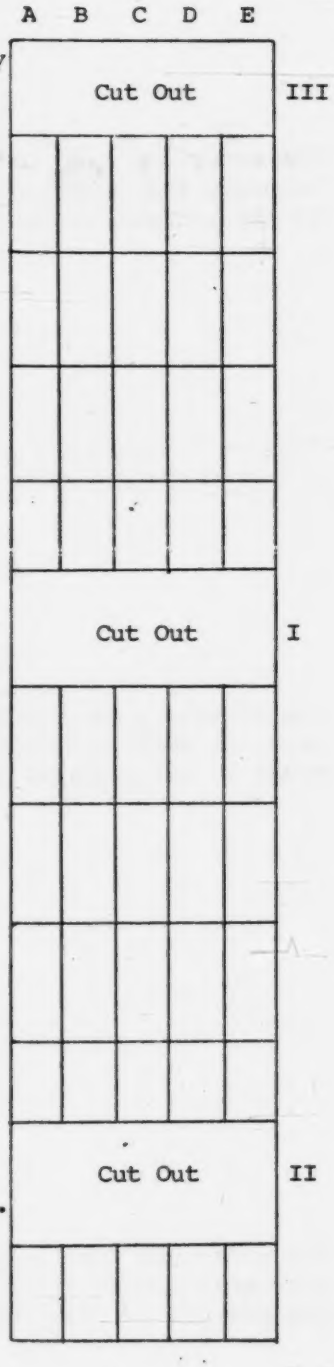


Scoring Key 5. Learning Achieved

1. I learned very little from this activity and I don't think it has helped me prepare for Volunteer service.

6. I learned a great deal from this activity which I feel has helped me prepare for Volunteer service.

11. I learned a moderate amount from this activity and some of what I learned has helped me prepare for Volunteer service.



TRAINING ACTIVITY EVALUATION SCALES

Scoring Matrix

Descriptive Statements			Points
I (High)	II (Medium)	III (Low)	
+	+	+	7
+	+	0	7
+	+	-	7
0	+	+	6
0	+	0	6
0	+	-	6
-	+	+	5
-	+	0	5
-	+	-	5
0	-	+	5
-	0	+	4
0	0	-	4
+	0	+	4
+	0	-	4
0	0	0	4
-	-	+	3
-	0	-	3
+	-	+	3
+	0	0	3
-	-	0	2
0	-	0	2
+	-	0	2
-	-	-	1
+	-	-	1
0	-	-	1

+ = The activity is <u>better than</u> the statement.
0 = The statement <u>matches</u> the activity.
- = The activity is <u>worse than</u> the statement.

TRAINING ACTIVITY EVALUATION SCALES

Scoring Work Sheets

Activity A

Scales → / Trainees ↓	1. Clarity of Objective	2. Skill of Trainer	3. Effectiveness of Method	4. Quality of Material	5. Learning Achieved
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
Sums	1.	2.	3.	4.	5.
Means $\left(\frac{\text{Sum}}{N}\right)$	1. ↓	2. ↓	3. ↓	4. ↓	5. ↓

Write these values down column A on the summary sheet.

Activity B

Scales → / Trainees ↓	1. Clarity of Objective	2. Skill of Trainer	3. Effectiveness of Method	4. Quality of Material	5. Learning Achieved
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
Sums	1.	2.	3.	4.	5.
Means $\left(\frac{\text{Sum}}{N}\right)$	1. ↓	2. ↓	3. ↓	4. ↓	5. ↓

Write these values down column A on the summary sheet.

Activity C

Scales → / Trainees ↓	1. Clarity of Objective	2. Skill of Trainer	3. Effectiveness of Method	4. Quality of Material	5. Learning Achieved
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
Sums	1.	2.	3.	4.	5.
Means $\left(\frac{\text{Sum}}{N}\right)$	1. ↓	2. ↓	3. ↓	4. ↓	5. ↓

Write these values down column A on the summary sheet.

Activity D

Scales → / Trainees ↓	1. Clarity of Objective	2. Skill of Trainer	3. Effectiveness of Method	4. Quality of Material	5. Learning Achieved
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
Sums	1.	2.	3.	4.	5.
Means $\left(\frac{\text{Sum}}{N}\right)$	1. ↓	2. ↓	3. ↓	4. ↓	5. ↓

Write these values down column A on the summary sheet.

Activity E

Scales → / Trainees ↓	1. Clarity of Objective	2. Skill of Trainer	3. Effectiveness of Method	4. Quality of Material	5. Learning Achieved
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
Sums	1.	2.	3.	4.	5.
Means $\left(\frac{\text{Sum}}{N}\right)$	1. ↓	2. ↓	3. ↓	4. ↓	5. ↓

Write these values down column A on the summary sheet.

TRAINING ACTIVITY EVALUATION SCALES

Summary Sheet

Training Program _____

Week Evaluated _____

Date _____

	Activity A	Activity B	Activity C	Activity D	Activity E
Scale 1. Clarity of Objective					
Scale 2. Skill of Trainer					
Scale 3. Effectiveness of Method					
Scale 4. Quality of Materials					
Scale 5. Learning Achieved					

Summary of Comments and Suggestions

Activity Evaluated: _____

TRAINING ACTIVITY EVALUATION SCALES

Error Computation Work Sheet

Scales → / Trainees	1. Clarity of Objective	Rater Errors	2. Skill of Trainer	Rater Errors	3. Effectiveness of Method	Rater Errors	4. Quality of Materials	Rater Errors	5. Learning Achieved	Rater Errors	Rater Error Sums
1											__ out of 5
2											__ out of 5
3											__ out of 5
4											__ out of 5
5											__ out of 5
6											__ out of 5
7											__ out of 5
8											__ out of 5
9											__ out of 5
10											__ out of 5
11											__ out of 5
12											__ out of 5
13											__ out of 5
14											__ out of 5
15											__ out of 5
16											__ out of 5
17											__ out of 5
18											__ out of 5
19											__ out of 5
20											__ out of 5
Scale Error Sums =		____ out of 20		____ out of 20		____ out of 20		____ out of 20		____ out of 20	Average Rater Error Sum/N = __ out of 5
Scale Error % Sum/20		1. __%		2. __%		3. __%		4. __%		5. __%	Rater Error % = __% Aver. Sum/5

ERROR KEY

I (High)	II (Medium)	III (Low)
+	+	+
0	+	+
-	+	+
-	0	+
-	-	+
-	-	0
-	-	-

\+ = The activity or program is <u>better than</u> the statement.
0 = The statement <u>matches</u> the activity or program.
\- = The activity or program is <u>worse than</u> the statement.

USE OF KEY: Any set of ratings other than the seven listed here is in error.

IV. CONCLUSION

The two evaluation scaling systems produced during this project represent significant improvements over conventional methods. Their systematic use in evaluating Peace Corps training would result in much more accurate and reliable information than the methods now being employed. The weaknesses in the present system that these two methods will overcome, as well as other benefits of the new scaling procedures, have been discussed in previous sections. The new methods do have drawbacks, however, two of which are listed below along with recommendations for modification:

- The nine criterion categories, along with the twenty-seven degree-of-effectiveness statements in the Training Program Evaluation Scales, represent what the participants (former trainees) in this study thought were the most important aspects of training. These may not be adequate for particular purposes of evaluation inquiry, or for other training programs. It is recommended that where these categories seem inappropriate, new and more relevant ones be developed according to the Retranslation procedure.

- Compared with conventional rating scales, completing these scales is rather a laborious task. (For the Training Program Evaluation Scales, the rater must make thirty-two separate judgments; for the Training Activity Evaluation Scales, the rater must make seventy-five different judgments in assessing five different training activities.) Furthermore, after the statements have been checked by the rater, the scoring is somewhat time consuming.

  The extra effort that this complexity and sophistication represents seems worthwhile in light of the consequences of decisions made based on evaluation data. More accurate and reliable data will require more time and effort. However, where this system

seems too complicated, a modification can be made that greatly simplifies the system, although the accuracy and reliability of the data suffers. This modification involves eliminating the use of the Mixed Standard Scaling method and replacing it with a simple three- or five-point continuous scale for each criterion category. For example, the three statements describing high, medium, and low effectiveness for the "Quality of Training Materials" category would be arranged along a five-point scale. The rater would make one judgment (instead of three) for this rating by selecting a number from one to five.

In conclusion, it is recommended that these new scaling systems be put to use in evaluating Peace Corps training programs, and that data be collected on a systematic basis in order to render the systems maximally useful. The rater, scale, and activity reliabilities should be computed and recorded, so that decisions can be made with known degrees of confidence and so that the scales themselves can be modified for greater usefulness, accuracy, and reliability.

REFERENCES

Barrett, R. S., Performance Rating. Chicago: Science Research Association, Inc., 1966.

Bass, B. M., "Further Evidence on the Dynamic Character of Criteria," Personnel Psychology, Vol. 15, 1962.

Beatty, Walcott H., (Chairman and Editor), Improving Educational Assessment and an Inventory of Measures of Affective Behavior. Association for Supervision and Curriculum Development, NEA, Washington, D. C., 1969.

Bend, Emil, "The Impact of the Social Setting upon Evaluative Research," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Blanz, Friedrich and Edwin E. Ghiselli, "The Mixed Standard Scale: A New Rating System," Personnel Psychology, Vol. 25, No. 2, 1972.

Bloom, Benjamin S., J. Thomas Hasting, and George F. Madaus, Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Campbell, Donald T., and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research in Teaching," in N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Carver, Ronald P., "Special Problems in Measuring Change," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Center for Research and Education Training Directors, Final Reports, Brazil Peace Corps Training Programs, 1972-1973. Center for Research and Education, Denver, Colorado (Peace Corps Contract PC-72-42043).

Edwards, A. L., Techniques of Attitude Scale Construction. New York: Appleton-Century Crofts, 1957.

Fitzpatrick, Robert, "The Selection of Measures for Evaluating Programs," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Garner, W. R., "Rating Scales, Discriminability, and Information Transmission," Psychological Review, Vol. 67, 1960.

Glass, Gene V., The Growth of Evaluation Methodology, Research Paper No. 27, Laboratory of Educational Research, University of Colorado, March 1969.

Glennan, Jr., Thomas K., Evaluating Federal Manpower Programs: Notes and Observations. Rand Corporation, Santa Monica, California, September 1969.

Gold, Norman, "An Illustration: Evaluating a Complex Social Program," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Guion, R. M., "Criterion Measurement and Personnel Judgments," Personnel Psychology, Vol. 14, 1961.

Hawkridge, David G., "Designs for Evaluative Studies," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Johnson, George H., "The Purpose of Evaluation and the Role of the Evaluator," Evaluative Strategies and Methods, American Institutes of Research, 1970.

Kaufman, Roger A., Educational System Planning. Englewood Cliffs: Prentice-Hall, Inc., 1972

Margolis, Frederic , and Steve Gillespie, Programming for Volunteer Service (developed for ACTION). F. M. Associates Ltd., Rockville, Maryland.

Osgood, Charles E., "Exploration in Semantic Space: A Personal Diary," Journal of Social Issues, Vol. 27., No. 4, 1971.

Prien, E. P., "Dynamic Character of Criteria: Organization Change," Journal of Applied Psychology, Vol. 50, 1966.

Schultz, D. G., and A. I. Siegel, Post-Training Performance Criterion Development and Applications: A Selective Review of Methods for Measuring Individual Differences in On-the-job- Performance. Applied Psychology Service, Wayne, Pennsylvania, 1961.

Scrivens, Michael, "Methodology of Evaluation," Perspectives of Curriculum Evaluation, Vol. 1. New York: Rand McNally and Company, 1967.

Smith, Patricia Cain, and L. M. Kendall, "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales," Journal of Applied Psychology, Vol. 47, No. 2, 1963.

Stake, Robert E., "Language, Rationality, and Assessment," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, Association for Supervision and Curriculum Development, NEA, 1969.

Stevick, Earl, Memory, Meaning and Method: Some Psychological Perspectives for Language Teachers. Foreign Service Institute, Department of State, Washington, D. C. (In draft form, not yet published, 1973.)

Stufflebeam, Daniel L., "Evaluation as Enlightenment for Decision-Making," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, Association for Supervision and Curriculum Development, NEA, 1969.

Tatsuoka, Maurice M., Nationwide Evaluation and Experimental Design, University of Illinois at Urbana-Champaign, 1972. Paper prepared for the 1972 Annual Meeting of the American Educational Research Association.

Torgerson, Warren S., Theory and Methods of Scaling. New York: John Wiley & Sons, Inc., 1958.

Tyler, Ralph W., "The Purposes of Assessment," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, Association for Supervision and Curriculum Development, NEA, 1969.

Wallace, S. R., "Criteria for What?" American Psychologist, Vol. 20, 1965.

Wight, Albert R., William L. Wight, and Mary Anne Hammons, Guidelines for Peace Corps Cross-Cultural Training, Parts I, II, III, IV. Center for Research and Education, Denver, Colorado, March 1970 (Peace Corps Contract PC -25-1710).